PROJET 2

Application detectant des produits contenant l'huile de palme

LO Ousmane , IML

Openclassrooms

May 17, 2022

SOMMAIRE

1. Introduction
2. Nettoyage des données
3. Exploration des données
4. Conclusion

Introduction

Santé publique France

Jeu de données foot :

Nombre ligne	Nombre de colonne	Nombre de types		
		objet	flaot	int
320772	162	105	56	1

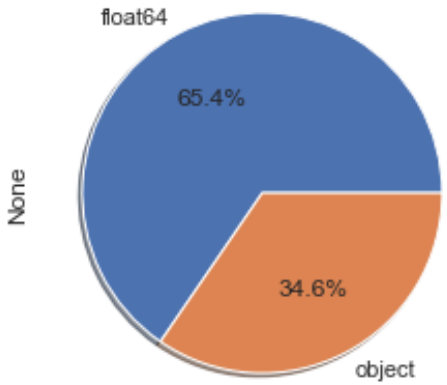


Figure: Piechart des types.

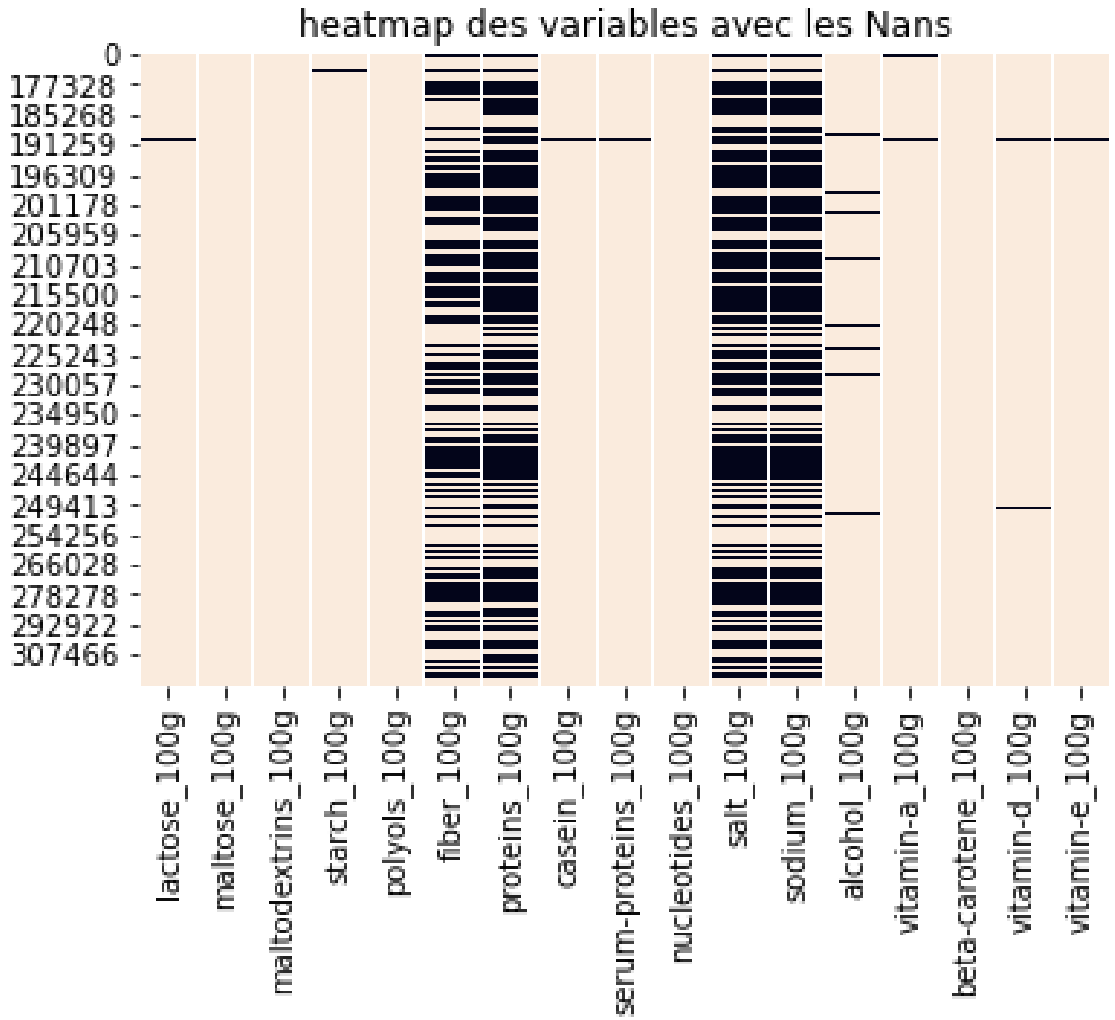
Objectif

Pour etudier la faisabilité de mon application je souhaite verifier:

- Les variables qualiTatives liées avec les ingreédients palm oil
- Les variables quantitatives liées avec les ingreédients palm oil
- les produits contenant de l'huile de palm
- la nutrition grade des produits qui en contiennt

Nettoyage des données

On observe plusieurs valeurs manquantes sur le jeu données.



Nettoyage des données

Suppression des variables contenant plus 80% de nans

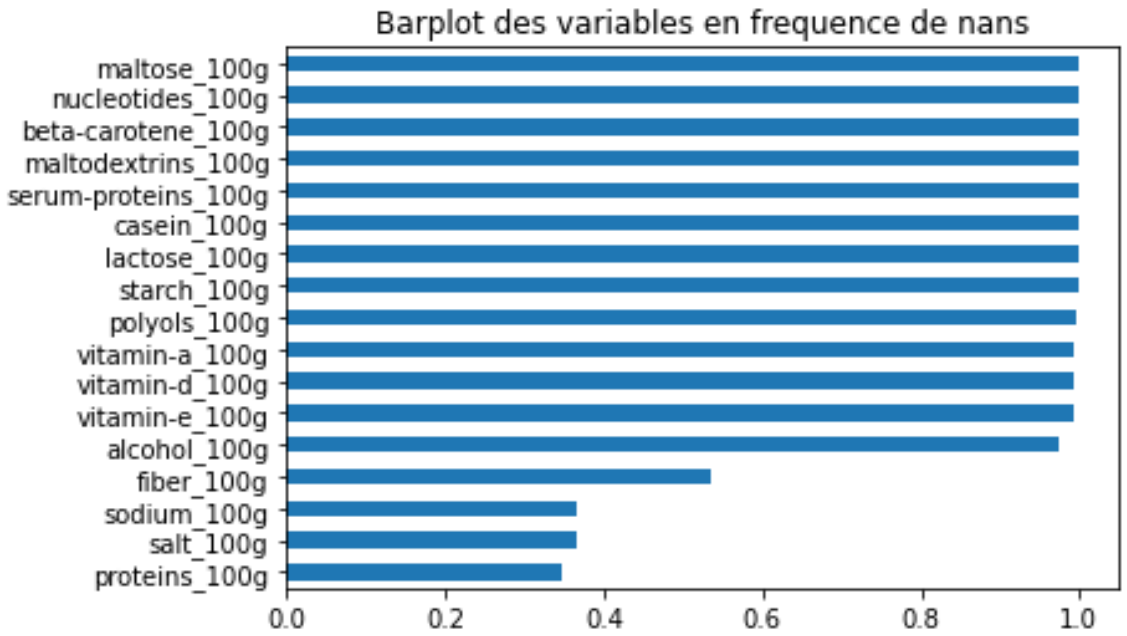
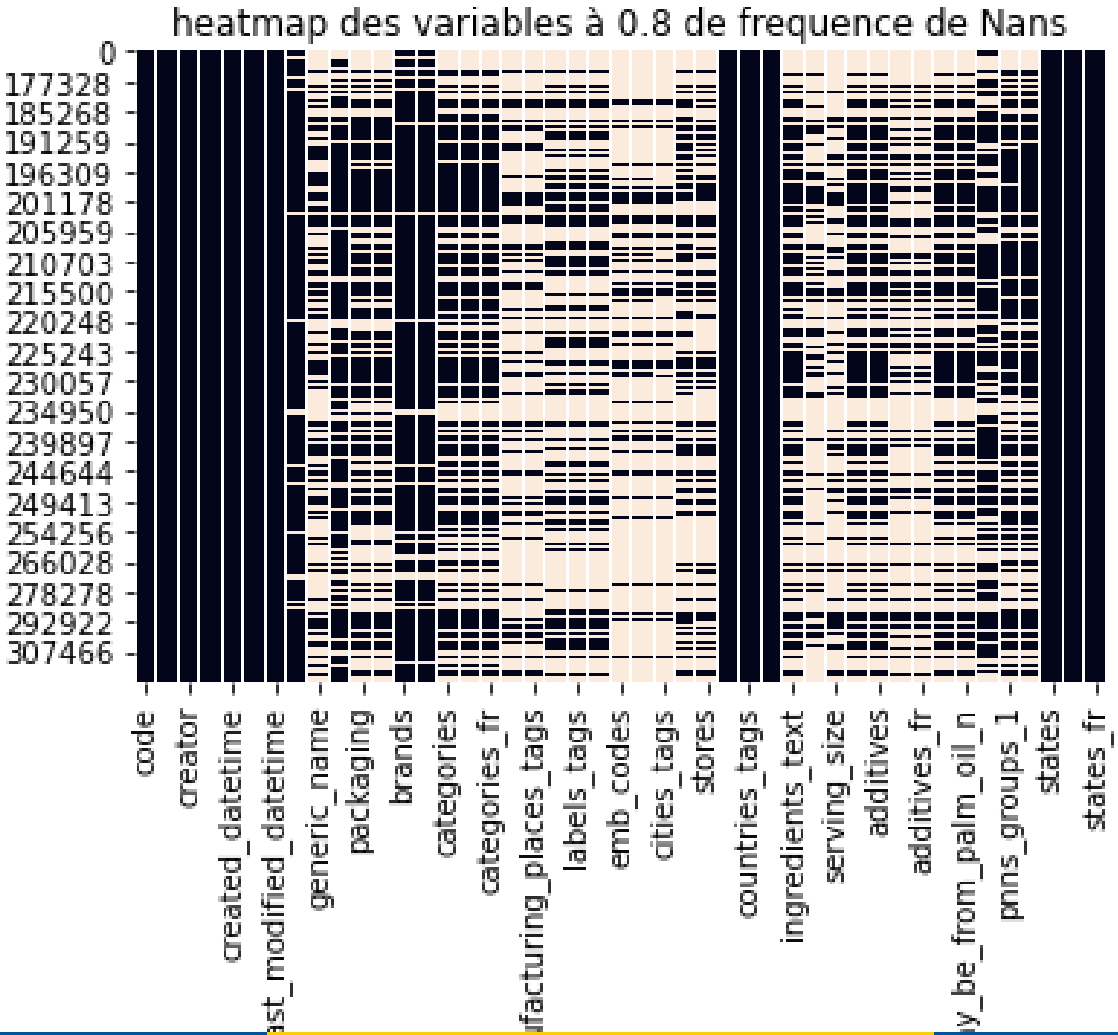


Figure: Barplot des nans par variable.

Nettoyage des données

Il existe maintenat moins de nans dans le jeux de données après suppréssion.



Doublons et Outliers

On a effectué ces actions avant de passer à l'imputation des nans:

- Détection des doublons
- Suppréssion des doublons
- Détections des outliers
- Remplacement des outliers par des nans

Imputation des Nans

On souhaite remplacer les nans soit par 0, la moyene ou la médiane
Mais en Gardant la variance ou l'information de 95 *à* 99%

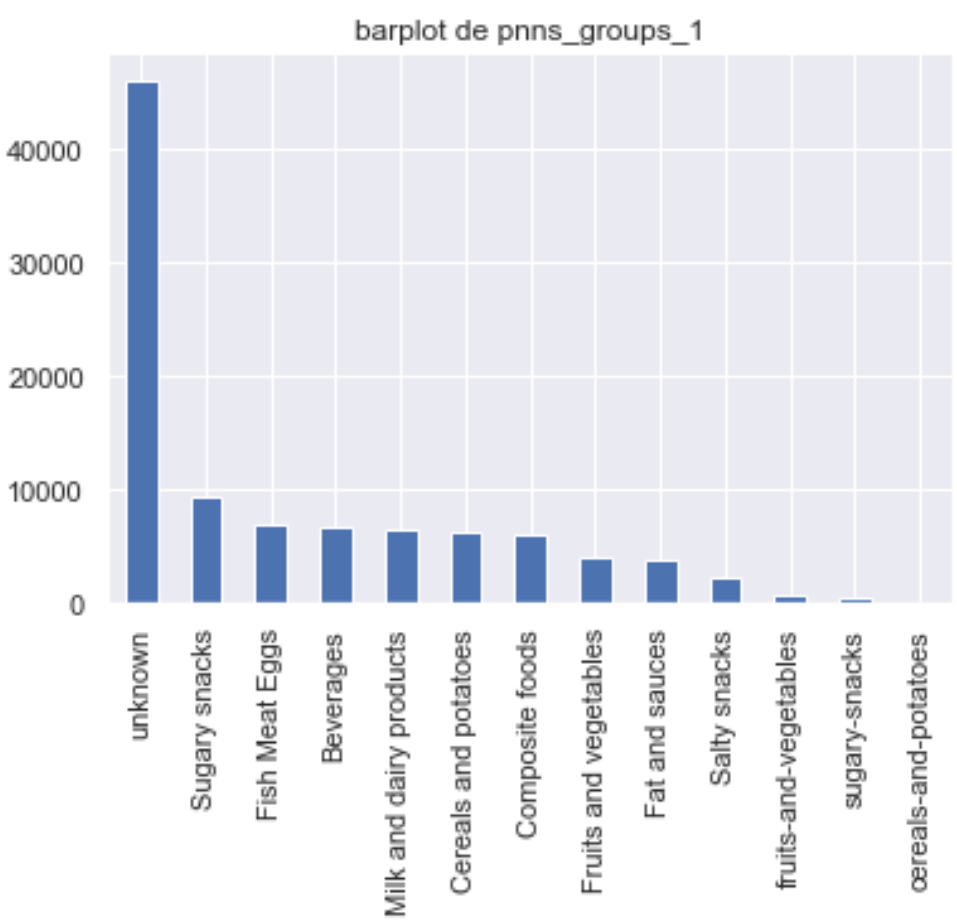
P 0	P mean	P median
100.05	99.95	99.94

Figure: Tableau Pourcentages des variances avec (0, mean, median)

Finalement les nans des données sont remplacés:

- Par le mode sur les variables catégorielles
- Par la médiane sur les variables continues

Exploration des données



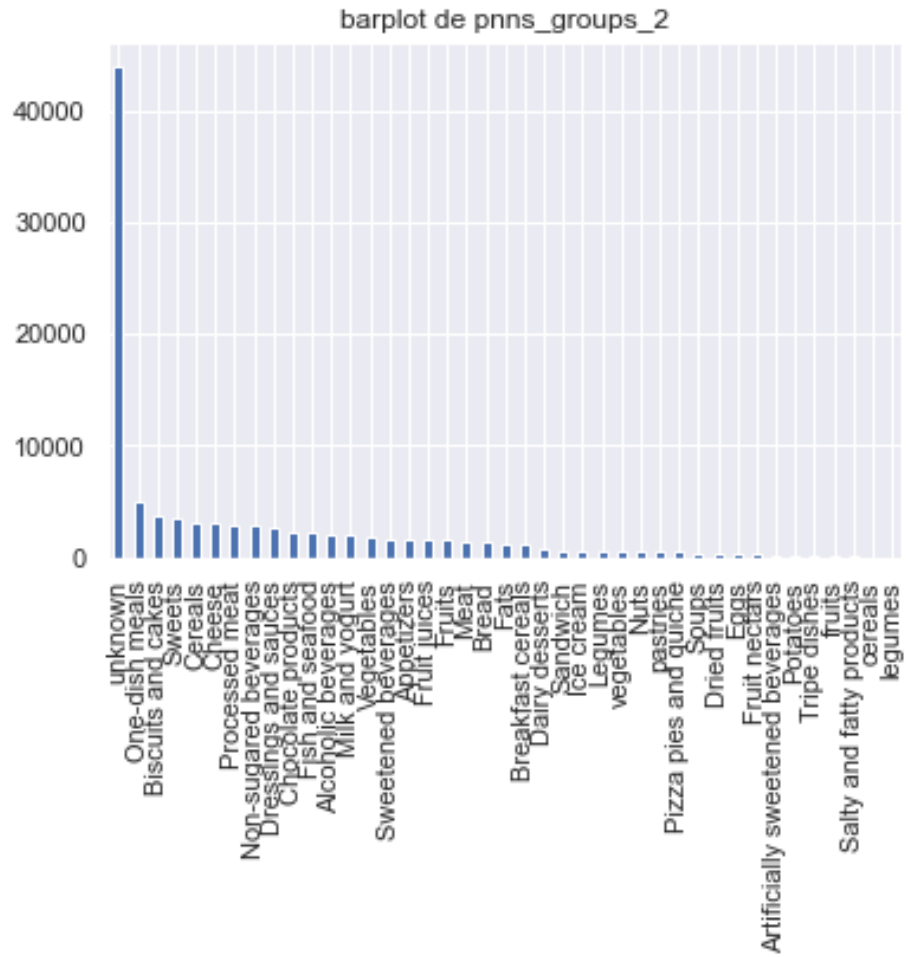


Figure: Barplot des variables pnns-groups

Exploration des données

Figure: Barplot de nutrition-grade-fr

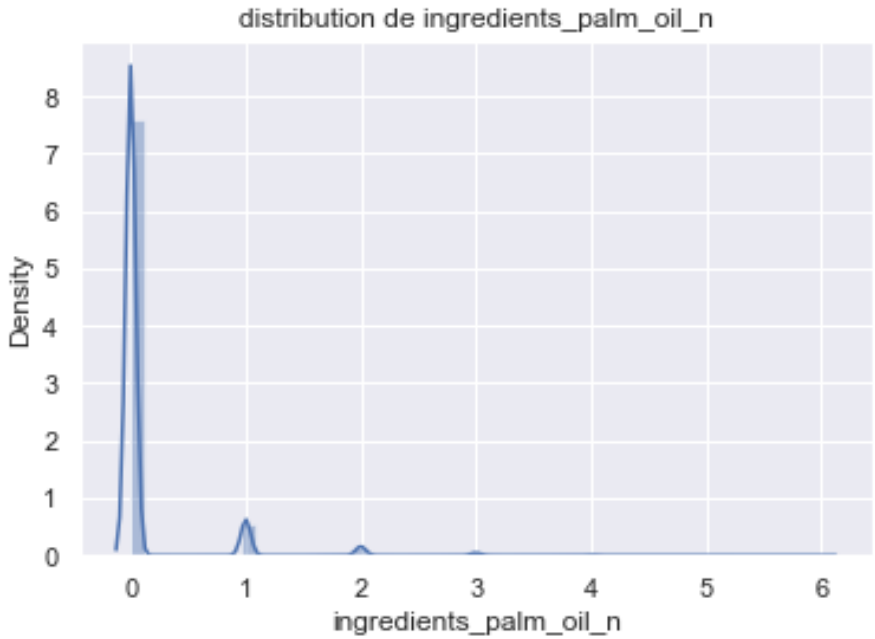


Figure: distribution d' ingredient-palm-oil-n

ACP

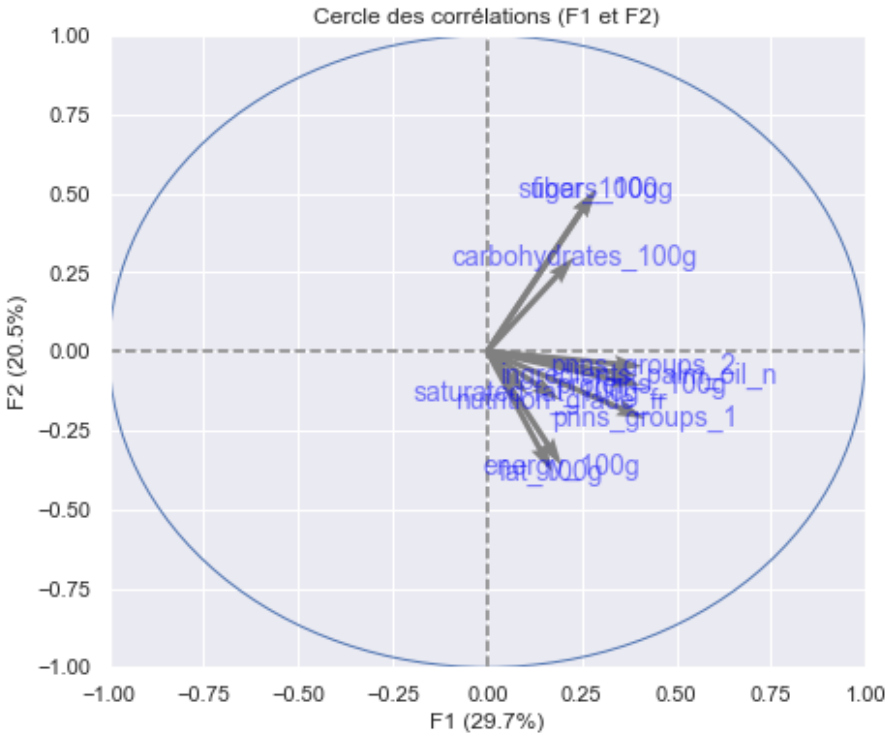


Figure: F1-F2

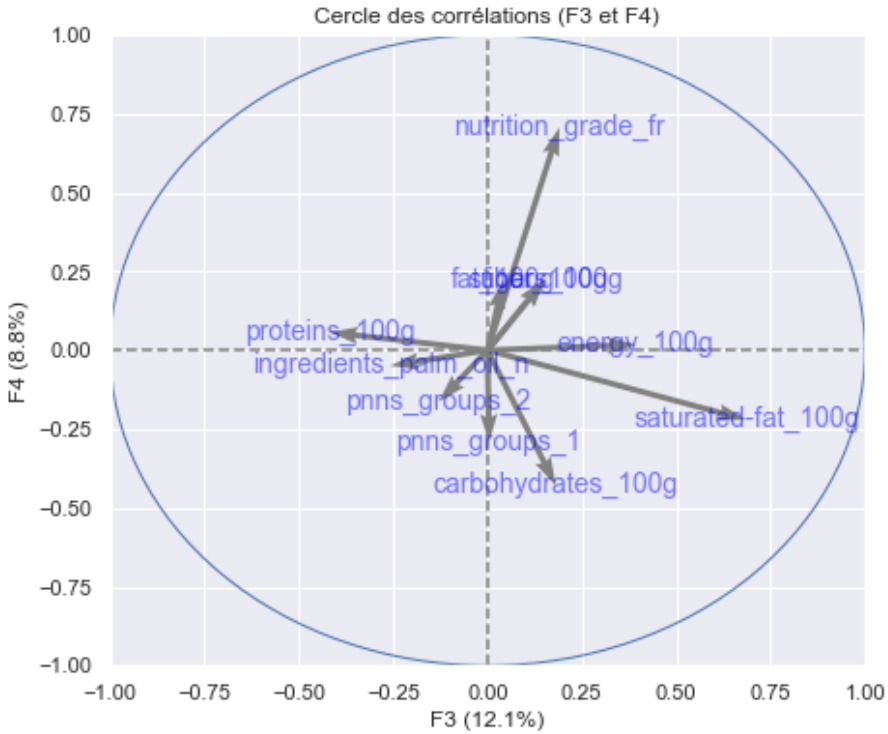


Figure: F3-F4

ACP

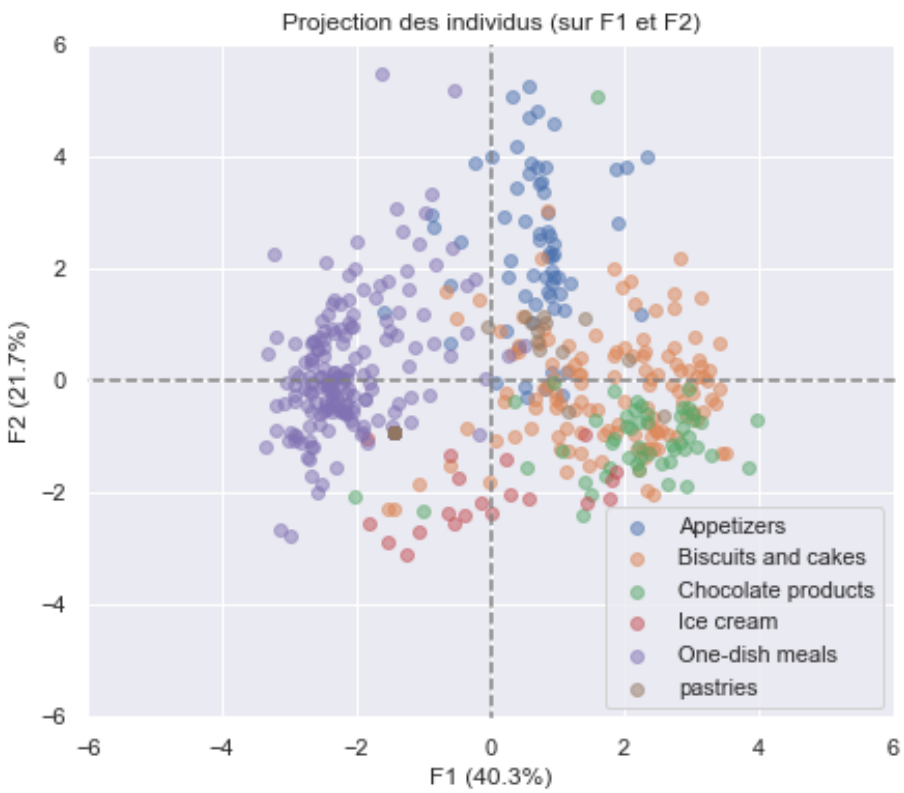


Figure: F1-F2

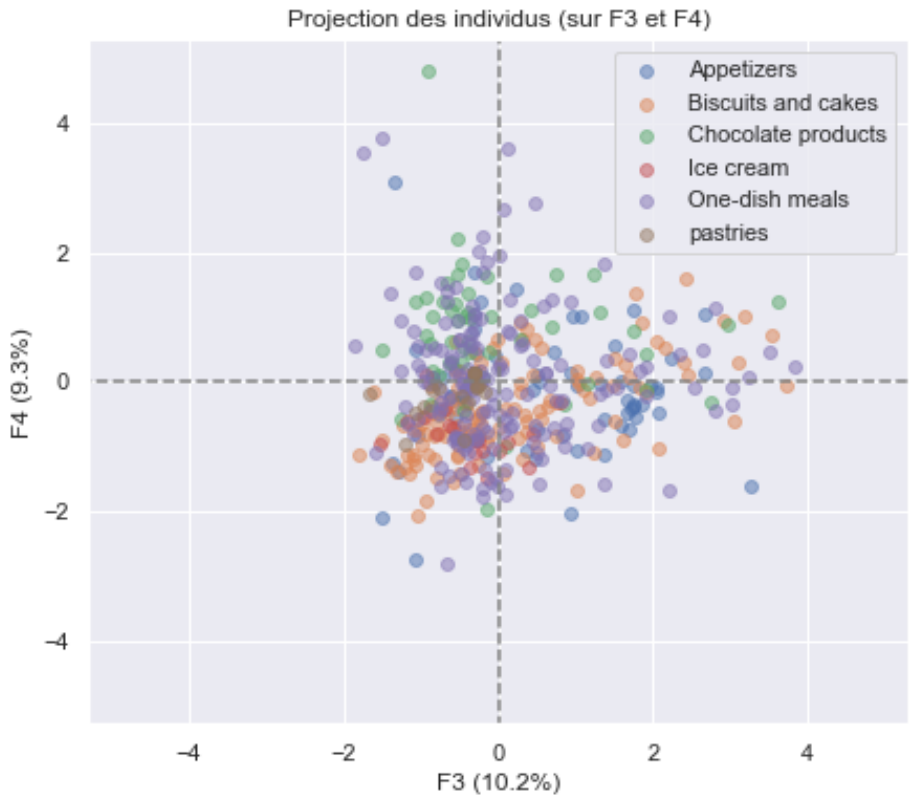
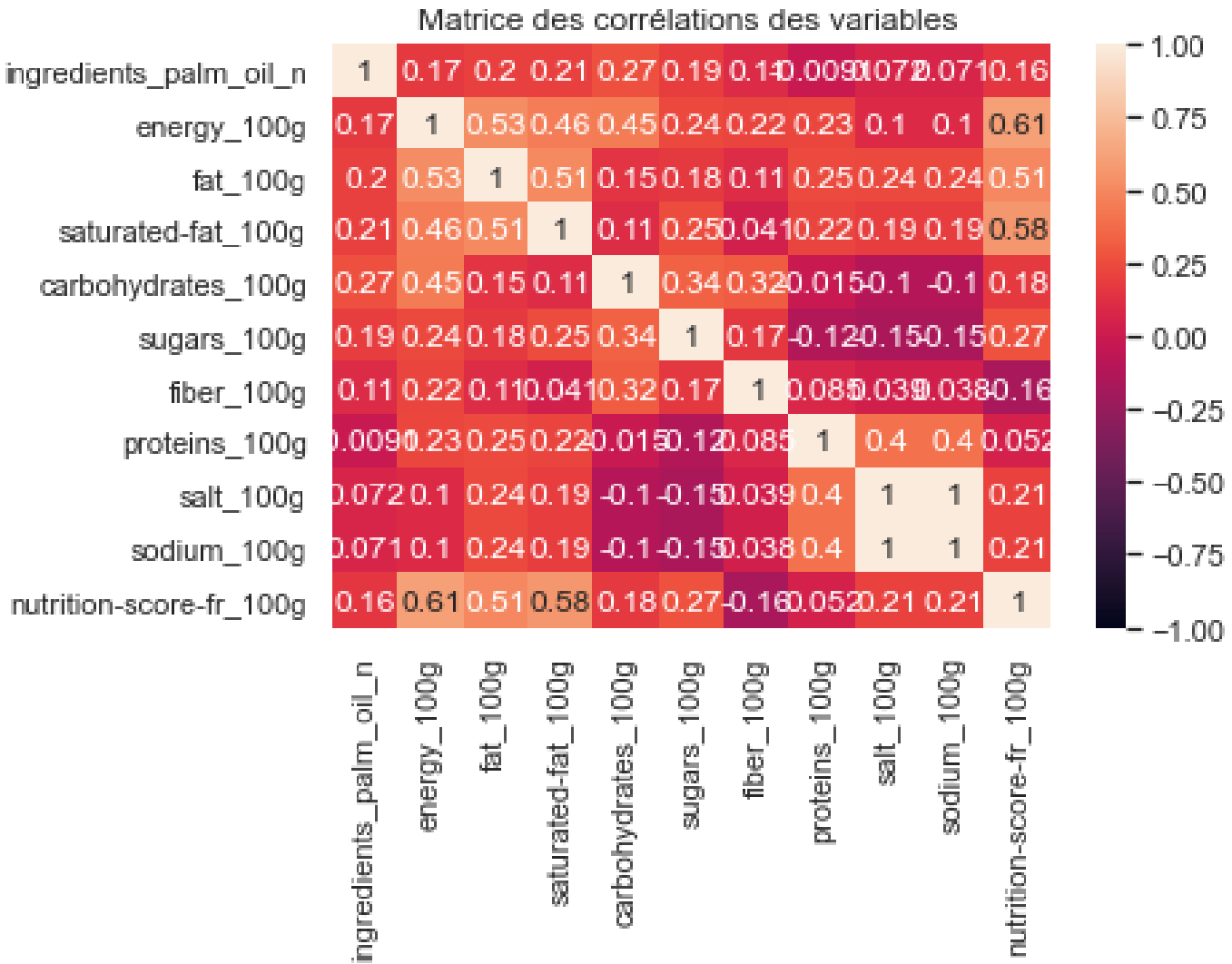


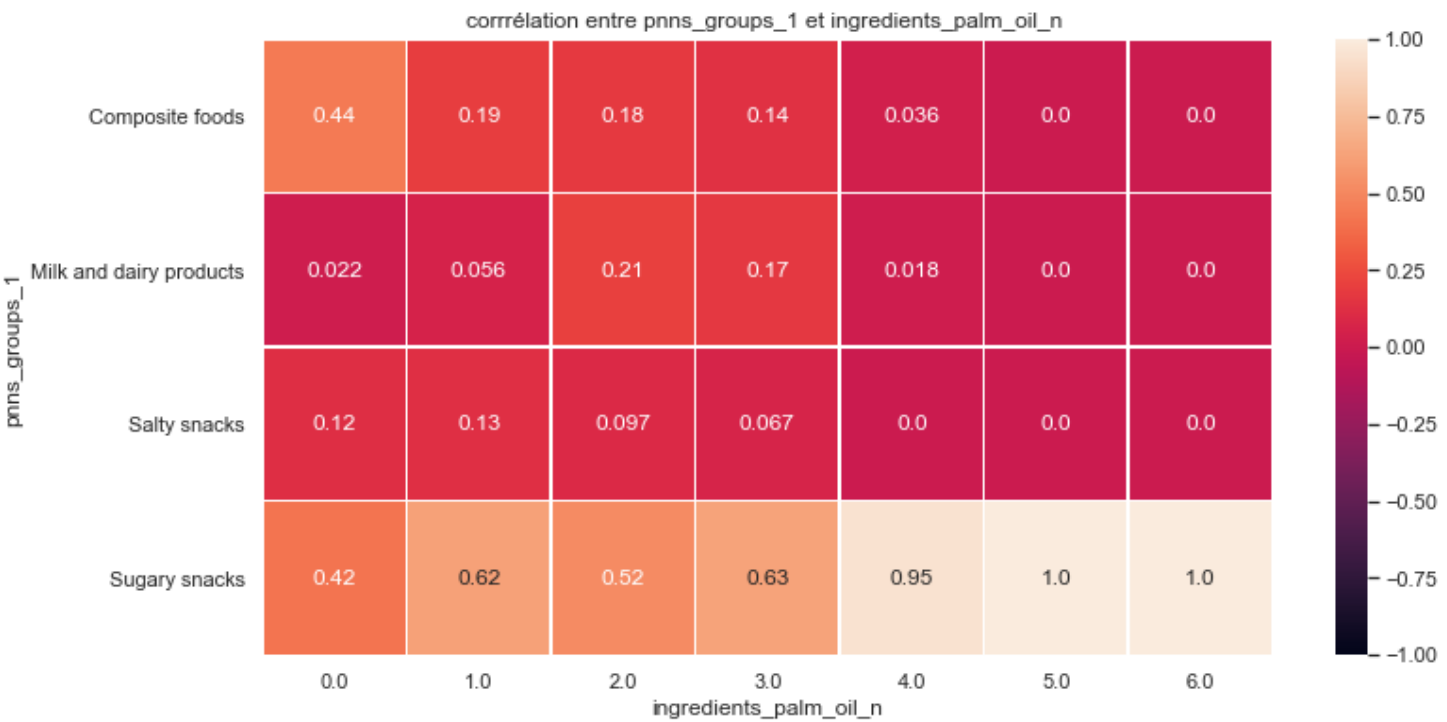
Figure: F3-F4

Corrélation des variables



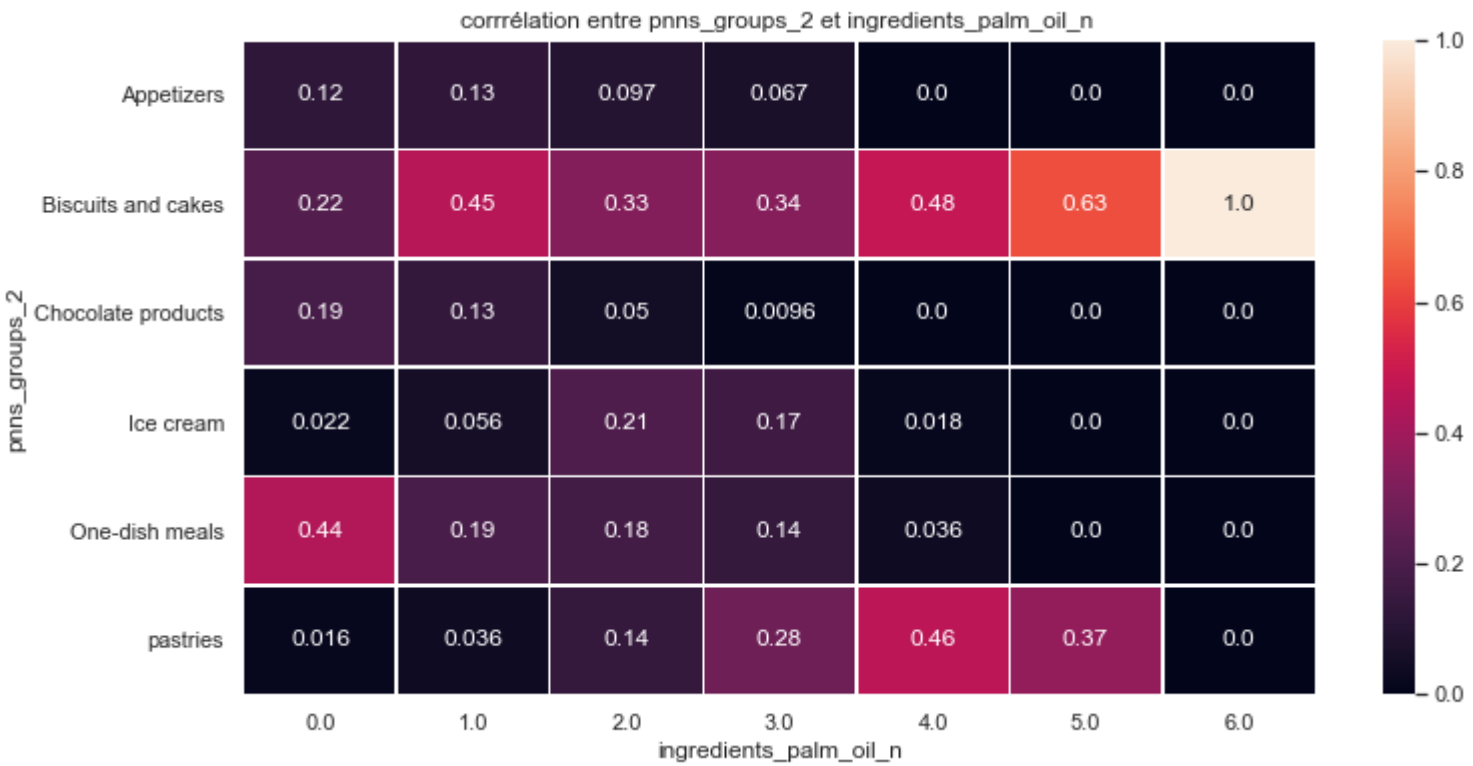
- Variables corrélées entre elles

Corrélation des variables



- les deux Variables sont liées d'après le heatmap

Corrélation des variables



- les deux Variables sont liées d'après le heatmap

Anova à un facteur

	sum_sq	df	F	PR(>F)
pnns_groups_1	2134.830501	12.0	1124.759722	0.0
Residual	15572.565770	98455.0	NaN	NaN

	sum_sq	df	F	PR(>F)
pnns_groups_1	2134.830501	12.0	1124.759722	0.0
Residual	15572.565770	98455.0	NaN	NaN

	sum_sq	df	F	PR(>F)
nutrition_grade_fr	290.832330	4.0	411.048697	0.0
Residual	17416.563941	98463.0	NaN	NaN

- p-values inferieurs à 0.05
- les variables sont liées à la variable ingrédient-palm-oil-n.

Anova à 2 facteurs

	sum_sq	df	F	PR(>F)
C(pnns_groups_1)	1433.967859	12.0	779.814843	0.000000e+00
C(nutrition_grade_fr)	31.612350	4.0	51.573916	4.106417e-23
C(pnns_groups_1):C(nutrition_grade_fr)	1240.194424	48.0	168.609431	0.000000e+00
Residual	15079.851999	98408.0	NaN	NaN

- p-values inferieurs à 0.05
- l'interaction a un effet au ingrédient-palm-oil-n.

Conclusion

L'analyse du jeux de données montre que :

- l'huile de palme est riche en en lipide et sucre
- certains produits alimentaires contiennnent un taux important d'ingredient de l'huile de palme
- l'huile de palme est souvent utilisé sur certains produits insdustriels tels ques les biscuits, les chocolats etc
- plus qu'un produit contient de l'huile de palme plus sa nutrition est mauvaise